GIR'05
2005 ACM Workshop on Geographical Information Retrieval

Christopher B. Jones
School of Computer Science
Cardiff University
c.b.jones@cs.cf.ac.uk

Ross Purves
Department of Geography
University of Zurich
rsp@geo.unizh.ch

1 Introduction

Geographical information retrieval (GIR) is concerned with the problems of finding information resources that relate to particular geographical locations. Until recently most web search engines have treated geographical terminology within user queries in the same way as other terminology. This can often result in failure to find relevant documents and in the retrieval of irrelevant documents. There are several reasons for this. For example, there are many different places with the same name, so unless the query contains a unique set of geographical terms, documents referring to the wrong place may be retrieved. There are many uses of place names in the names of people and organisations with the consequence that search depending on naive exact matching of terms may retrieve such documents even though they may not relate to the named place. It is also the case that queries may include spatial prepositions, such as near or outside, which will not be interpreted intelligently by the search engine. Thus relevant documents may not be found because the user's query did not contain explicit references to the places that match the query expression, in the sense of being, for example, near, inside or north of a named place. It is also the case that conventional search engines lack other desirable facilities for geographical search such as relevance ranking that takes account of geographical proximity and the ability to use interactive maps to specify a query and to visualise the results.

Limitations such as these in conventional search techniques have led to a growing body of research concerned with various aspects of geographical information retrieval. This includes topics such as geo-parsing to detect geographical terminology in documents, geo-coding to attach an unique locational reference to a document, as well as the subjects of spatial indexing of documents, geographical relevance ranking, user interfaces for geographical retrieval, retrieval of geo-data for use in geographical information systems and issues concerned with the design and evaluation of GIR systems,

The Workshop on Geographical Information Retrieval at CIKM'05 in Bremen, Germany in November 2005 was the successor to the very successful first GIR Workshop which was held in conjunction with SIGIR 2004 in Sheffield, UK. Both events attracted a mix of delegates with backgrounds in information retrieval and in geographical information science. Details of the programme for GIR'05 can be found at http://www.geo.unizh.ch/~rsp/gir-cikm/. We now summarise the content of the twelve presentations, following the order of the workshop programme.

2 Summary of presentations

The first three papers addressed the subject of visualisation in GIR. The paper by Carmo, Freitas, Afonso and Claudio described a method for modifying the amount of retrieved information that is presented to a user in a way that adapts to the user's expected interests. A degree of interest function is based on attributes of the query topic and the distance of the displayed feature from the centre of the display. This is combined with multiple levels of detail of a map display. The approach is analogous to the cartographic process of generalisation. In future work it is envisaged that the user will be given control of specifying the attributes that determine the degree of interest.

Hobona, James and Fairburn reported on a multidimensional visual interface for GIR. They compared a 3-D graphics tool for visualising the results of information retrieval tasks with the approach of conventional ranked lists, as used in a different geo-data search facility. They found that their 3D visualisation method improved to some degree on the precision of results when compared with a ranked list, while also being found to be superior with regard to several qualitative measures of ease of use. The study indicates that users are becoming more accepting of 3D visualisation methods than has previously been the case, possibly as a result of greater familiarity with computer games in which 3D graphics are commonplace.

The third of these papers, by Albertoni, Bertone and De Martino, addressed the problem of mismatch between the user's understanding of a subject area and the data content and vocabulary of a particular set of resources that may be of interest. They proposed a query interface that allows the user to visualise the data space in terms of specific categories of interest, individual values of which they may select to refine their search. This is combined with visualisation of an ontology representing the vocabulary of the subject, enabling the user to consider using alternative data specific terms that are close in meaning to their original query terms.

The identification of the geographical scope to which a web page refers is a major issue in GIR. Two papers in the workshop addressed this task from differing, but complimentary, perspectives. The work of Wang, Xie, Wang, Lu and Ma described three different forms of geographic location that can be defined through the use of web pages and related data. These are the physical location of the provider owning the web content, the content location - that is to say the place described by the content, and finally the serving location which is defined as the region within which the content is of interest. Both provider and content location are calculated through geoparsing of documents for references to geographic places found in gazetteers. A set of rules are used to distinguish between locations which are likely to refer to provider or content location. Thus, for example, it is noted that provider location is likely to appear on multiple pages and be found more often in the header or the footer of a page. Having identified possible candidates for provider location, a Support Vector Machine is used to extract a probable provider location.

When specifying the content location of a web page, a key challenge lies in dealing with the fact that many web pages refer to multiple locations. Wang et al. describe an approach to identifying dominant locations through the use of a weighted geographic tree based on the associations between the locations grounded within the web page. Finally, serving location is calculated using a similar algorithm, but using user locations (gleaned from server logs) and inbound hyperlink locations as additional data sources. Initial results from the application of these algorithms to a relatively small dataset show that, firstly, distinguishing between these different sorts of location is profitable since 76% of content locations are different from provider locations. Furthermore, for the test data used the algorithms appear to give good results in terms of both precision and recall.

Clough's paper is also concerned with the extraction of geographic location from web pages, though in this case for a considerably larger dataset. Clough's approach to the problem employs similar geographic data sources, but focuses on correctly identifying geographic content, of any sort, contained within the web pages. He employs techniques from Named Entity Recognition to geoparse the documents. Having grounded place names, Clough deals with the problem of referent ambiguity by the use of default senses. Thus, he assumes that the mostly likely location referred to by a placename is the one with the shortest administrative hierarchy (e.g. County -> Region -> Town). The approach is simple enough to be applied to a range of different European countries, since the rules applied are language independent. Clough reports on the performance of his approach for, once again, a relatively small sample data set. Finally, for four European countries, an analysis of the unique places identified in a collection of in excess of 800,000 web pages is presented, finding an average of around 3 geographic references per page.

Martins, Silva and Andrade reviewed different approaches to indexing and ranking in GIR as part of the route to developing a spatially aware Portuguese geographic search engine. The paper sets out classic text indexing methods before going on to examine the issues that must be considered when retrieving information on the basis of not only context, but geographical scope. The nature of similarity in terms of geographic scopes is examined, with relationships such as overlap, adjacency and containment between scopes introduced. Such relationships also require that we develop appropriate indexing and ranking methods, and Martins et al. list a number of possible spatial indexing methods before discussing whether an index which combines geographic and contextual scopes or separate indices is the best approach. Finally, they point out the importance of developing an evaluation methodology for GIR, a point taken up in a subsequent paper presented at the workshop.

The goal of the work by Desai, Knoblock, Chiang, Desai and Chen is to enable the retrieval of street maps for a specified place from image and mapping resources on the web. As part of a process of indexing the maps, methods have been developed to distinguish street maps from other types of image and then to categorise the street maps as either dense or sparse. Machine learning techniques are used to detect images of street maps. Other pattern recognition techniques are used to detect road network intersections, prior to employing map conflation methods to compare the map images with vector data in order to derive the map coordinates and scales of the street images.

Much of the workshop dealt with the issue of retrieving unstructured data which contain some geographic reference. However, there are large quantities of relatively structured spatial data (e.g. remote sensing imagery, terrain models or maps of specific geographic attributes), containing explicit georeferences, that are often generated as a consequence of research projects. Though such data often have considerable potential for reuse, they are sadly underutilised due to their lack of accessibility. Tschirner and Zipf described a portal based on open standards allowing such data to be exchanged and, crucially, located. At the core of the portal lies the acquisition of standards-based metadata. Through the use of a range of Open Geospatial Consortium standards these metadata can be queried textually and graphically through a map interface before the datasets are uploaded by interested users.

Another problem associated with the use of structured geographical information is taken up in the paper by Lutz. This is concerned with the problem of finding and composing geoprocessing services on the web in a manner that effectively matches the user's requirements to web services that are available. The example that he gives is that of services to measure the distance between two locations, given a service to provide the locations of airports. Issues to be addressed include the fact that there are different types of coordinate systems and different ways of measuring distance (e.g. in

2D, 3D, on a plane or on a surface). Ontologies are used to provide a common vocabulary for advertising and searching for service functions and associated data types along with pre-and post-conditions (such as types of input and output coordinates) of the functions, represented in first order logic (FOL). The FOL aspect enables inexact matching between pre and post conditions using subsumption (generalisation) relationships that may differ for the pre and post conditions. A prototype of the "matchmaking process" has been implemented but further work is still to be done on creating a web catalogue client to manage the process of registering and processing requests for services.

The presentation by Exposto, Macedo, Pina, Alves and Rufino focused on the issue of crawling the content of the web itself. They postulate that distributing crawlers amongst geographic zones will lead to improvements in efficiency, and that geographical scope of web pages is likely to be correlated with the geographic position of servers. They used a number of simple heuristics to attempt to assign zones to web servers, but found that only a few of these (the use of institution names in host servers, and the NetGeo service) located a significant number of server IPs. Initial evaluation of the results of distributed crawls based on geographically distributed partitioning showed decreased download times with respect to both centralised and site-hash based partitioning. The authors suggest that by taking into account router locations and network topology, as well as the heuristics described in the paper, they may be able to improve the reliability of locations assigned to web servers.

Delboni, Borges and Laender were concerned with automated interpretation of natural language query expressions for purposes of geographical web search. They propose the use of textual query expansion whereby a geographic query of the form <subject in-proximity-to landmark> is transformed to a disjunction of queries employing a variety of spatial relationships that refer to proximity, for example "near to", "in front of", "in the vicinity of " etc. An experiment using the Google search engine showed that query expansion of this type improved both recall and precision when compared with a single query phrase. Their intention is to enhance this form of query by attaching quantitative interpretations of the distance associated with the natural language expression and hence introduce a spatial dimension to the associated relevance ranking.

The final presentation of the workshop by Martins, Silva and Chaves, considered the need for systematic methods for evaluation of geographical information retrieval systems. The authors distinguish between several aspects of GIR to which evaluation may be addressed, in particular those of finding geo-references within documents, assigning a main geographic context to a document (geographic scope), geographical relevance ranking, user interaction and computational performance issues. The paper reviews some of the existing work on the various aspects of GIR evaluation and presents some initial results from the authors on the recognition and grounding (geocoding) of georeferences in text and of the assignment of geographical scopes to documents. The authors highlight the difficulties in comparing the results of existing evaluation efforts, caused by the current lack of a substantial standard document corpus that is marked up with respect to geographical context.